Economics Letters 124 (2014) 290–295

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

A comparison of city size distributions for China and India from 1950 to 2010

Jeff Luckstead [a,*], Stephen Devadoss [b]

[a] University of Arkansas, United States

[b] University of Idaho, United States

HIGHLIGHTS

- We analyze the size distribution of Chinese and Indian cities for 1950–2010.
- We consider lognormal, Pareto, and general Pareto distributions.
- Lognormal characterizes both country's city size distribution in the early periods.
- Pareto represents the Chinese city size distribution in 2010.
- Indian size distribution in 2000 and 2010 follows Zipf.

ARTICLE INFO

Article history:
Received 25 April 2014
Received in revised form
3 June 2014
Accepted 6 June 2014
Available online 14 June 2014

JEL classification:
D30
R12
C24
C46

Keywords:
China
City size
General Pareto
India
Lognormal
Pareto

ABSTRACT

We examine the distributions of Chinese and Indian city sizes for seven decades (1950s to 2010s) using lognormal, Pareto, and general Pareto distributions. We ascertain which distribution fits the data and how the city size distributions change during these periods. The Chinese city size distribution is represented by lognormal in the early periods (1950–1990) and by Pareto in 2010, but is not characterized by Zipf, which could be attributed to Chinese government's restrictions of migration from rural to urban areas and the one-child policy. In contrast, the Indian city size distribution transitions from lognormal in the earlier periods to Zipf in the later periods.

1. Introduction

Zipf (1949) observed that Indian city sizes, as far back as 1911, followed a power law distribution, i.e., that the sizes of larger cities are inversely proportional to their ranks. Since this early work, many studies have observed this empirical regularity spanning several countries and time periods (Rosen and Resnick, 1980; Ioannides and Overman, 2003; Anderson and Ge, 2005). In particular, studies that considered the 135 largest cities in the United States generally have shown that Zipf's law holds (Krugman, 1996; Gabaix, 1999).

In this study, we consider size distribution of cities in the two most populous countries: China and India. Specifically, we examine the distribution of upper-tail cities for every decade between 1950 and 2010. Our results show for these largest cities, Zipf's law does not hold for China for all decades; however, for the last two decades (2000 and 2010), the size distribution is close to, but not quite, Zipf. The reason for this result could be that since 1950 China restricted population mobility from rural to urban areas through the Hukou system, but relaxed these policies on a limited basis in recent decades after the economic reforms in 1978. Zipf's law also does not apply for India for the early decades (1950–1990) because

* Corresponding author. Tel.: +1 208 310 1864; fax: +1 479 575 5306.
E-mail address: jluckste@uark.edu (J. Luckstead).

http://dx.doi.org/10.1016/j.econlet.2014.06.002

of the predominance of rural population with less economic incentives to move to urban areas. However, for the recent two decades (2000 and 2010) Zipf's law holds because increased mobility of workers from rural to urban areas due to economic reforms.

2. Methodology

We use lognormal, Pareto, and general Pareto distributions to estimate city size distributions for China and India and highlight the distribution that fits the data best. We apply maximum likelihood to estimate the parameters of these distributions and ascertain the fit using the Kolmogorov–Smirnov (*KS*) test, mean squared error, and Zipf plots which graph log of rank in the vertical axis and log of the city size in the horizontal axis. We also employ the Lagrangian multiplier test developed by Urzúa (2000) to determine whether the city size follows Zipf's Law. Using these approaches, we evaluate the historical evolutions in the city size distributions for these two countries.

2.1. Lognormal distribution

The joint lognormal PDF[1] for n i.i.d. samples of x is

$$f^L(x_1,\ldots,x_n;\mu,\sigma)=\prod_{i=1}^{n}\frac{1}{x_i\sigma\,(2\pi)^{1/2}}\exp\left(-\frac{(\log x_i-\mu)^2}{2\sigma^2}\right),$$

and the joint log likelihood is

$$\mathcal{L}^L(\mu,\sigma|x_1,\ldots,x_n)=-\sum_{i=1}^{n}\log x_i-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log\left(\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\log x_i-\mu)^2.$$

Optimizing this function with respect to μ and σ, we obtain

$$\hat{\mu}=\frac{\sum_{i=1}^{n}\log x_i}{n}\quad\text{and}\quad\hat{\sigma}^2=\frac{\sum_{i=1}^{n}\left(\log x_i-\hat{\mu}\right)^2}{n},$$

which can be estimated using the sample data for x. By substituting these estimates into the inverted lognormal CDF $\hat{x}^L=\exp\left(\hat{\sigma}2^{1/2}\,\mathrm{erf}^{-1}(2F(x)-1)+\hat{\mu}\right)$, where $\mathrm{erf}(\cdot)$ is the error function, we can predict city sizes.

2.2. Pareto distribution

The joint Pareto PDF for n i.i.d. samples of x is

$$f^P(x_1,\ldots,x_{n-1};x_m,\alpha)=(\alpha)^{n-1}\left(x_m^{\alpha}\right)^{n-1}\prod_{i=1}^{n-1}\frac{1}{x_i^{\alpha+1}},$$

$$x_i\geq x_m,\ x_m>0,\ \alpha>0,$$

with the joint log-likelihood

$$\mathcal{L}^P(x_m,\alpha|x_1,\ldots,x_{n-1})=(n-1)\ln\alpha+(n-1)\,\alpha\ln x_m-(\alpha+1)\sum_{i=1}^{n-1}\ln(x_i).$$

Noting that $\hat{x}_m=\min(x)$, this function is optimized to obtain the Hill estimator of α,

$$\hat{\alpha}=\frac{n-1}{\sum_{i=1}^{n-1}\ln(x_i)-(n-1)\ln\hat{x}_m}.$$

Substitute the estimates $\hat{\alpha}$ and $\hat{x}_m$ into the inverted Pareto CDF $\hat{x}^P=\hat{x}_m\,(1-F(x))^{-1/\hat{\alpha}}$ to predict the city size. Observe that when $\alpha=1$, we get the familiar Zipf distribution.

2.3. General Pareto distribution

For n i.i.d. samples of x, the joint general Pareto density is

$$f^{GP}(x_1,\ldots,x_{n-1}|\phi,\theta,x_m)=\prod_{i=1}^{n-1}\frac{\phi}{\theta}\left(1+\frac{x_i-x_m}{\theta}\right)^{-(\phi+1)},$$

$$x_i\geq x_m,\ x_m>0,\ \text{and}\ \phi>0.$$

Note that when $\theta=x_m$, the general Pareto distribution becomes the Pareto distribution; thus the former nests the latter. The corresponding joint log-likelihood is

$$\mathcal{L}^{GP}(\phi,\theta,x_m|x_1,\ldots,x_{n-1})=(n-1)\ln(\phi)-(n-1)\ln(\theta)-(\phi+1)\sum_{i=1}^{n-1}\ln\left(1+\frac{x_i-x_m}{\theta}\right).$$

Since the optimization of this function does not yield an analytical solution, we numerically estimate the parameters $\hat{\phi}$ and $\hat{\theta}$. Substituting these estimates into the inverted general Pareto CDF $\hat{x}^{GP}=\hat{\theta}\,(1-F(x))^{-\frac{1}{\hat{\phi}}}+\hat{x}_m-\hat{\theta}$, we can predict the city sizes.

With $\theta=x_m$ and $\phi=1$, the general Pareto turns into the Zipf distribution. Consequently, we can test the null hypothesis $\theta=x_m$ and $\phi=1$ using the Lagrange multiplier (*LM*) test as highlighted by Urzúa (2000):

$$LM=4n\left[z_1^2+6z_1z_2+12z_2^2\right]\overset{a}{\sim}\chi_2^2$$

where $z_1=1-\frac{1}{n}\sum_{i=1}^{n}\ln\frac{x_i}{x_m}$ and $z_2=\frac{1}{2}-\frac{1}{n}\sum_{i=1}^{n}\frac{x_m}{x_i}$.

Finally, we use the predicted values ($\hat{x}^L$, $\hat{x}^P$, and $\hat{x}^{GP}$) from each of the above three distributions and compare them to actual values to ascertain the fit of the distributions using *KS* statistics, mean squared errors (*MSE*), and Zipf plots.

3. Analysis and results

We collected population of cities for China and India for each decade from 1950 to 2010 (United Nations, 2011). This data contains cities that had an urban agglomeration population of 750,000 inhabitants or more in 2011, and each decade has the same sample of cities. The number of cities for China is 142 and for India is 58.

Tables 1 and 2 present the estimated parameters, *KS* statistics, *MSE*s of the log of the actual and predicted values, and Lagrange multiplier test for China and India, respectively. Figs. 1 and 2 illustrate Zipf plots of actual and predicted values of the three distributions for the sample cities in these countries. For China, the mean of the lognormal distribution increases over the decades, indicating the population growth in cities. In contrast, the variance tends to decline over the period, implying the population differences among cities are narrowing, which indicates greater mobility of people in recent years, stemming from the economic reforms. Based on the *KS* statistics, the lognormal distribution statistically fits the data for the Chinese city sizes for the decades from 1950 to 1990, which are below the 5% critical value of 0.11. But, for the recent two decades (2000 and 2010), the lognormal distribution does not perform well. These results are also corroborated by the *MSE*s and Zipf plots (Fig. 1(a)–(g)). Our findings are consistent with the results reported by Anderson and Ge (2005).

The population dynamics and city size distribution in China can be attributed to Chinese government policies regarding mobility of workers. Since the 1950s, the Chinese government maintained a household registration record, known as Hukou (Wang, 2008).

[1] Stanley et al. (1995) have applied the lognormal to study the size distribution of firms.

Table 1
Distribution parameterization for China.

Year	Lognormal				Pareto				General Pareto					
	$\hat{\mu}$	$\hat{\sigma}^2$	KS[a]	MSE	$\hat{x}_{\min}$	$\hat{\alpha}^P$	KS[a]	MSE	$\hat{x}_{\min}$	$\hat{\phi}$	$\hat{\theta}$	KS[a]	MSE	LM[b]
1950	11.65	2.01	0.10	0.10	3 000	0.26	0.35	6.15	3 000	2.13	330,007	0.09	0.08	1424.38
1960	12.11	1.69	0.08	0.04	7 000	0.30	0.34	4.20	7 000	2.16	477,291	0.08	0.03	902.75
1970	12.47	1.27	0.06	0.01	17 000	0.36	0.32	2.69	17 000	2.18	616,071	0.09	0.03	445.20
1980	12.82	0.97	0.06	0.01	39 000	0.44	0.29	1.61	39 000	1.95	659,737	0.10	0.04	223.63
1990	13.23	0.75	0.07	0.01	76 000	0.50	0.29	1.24	76 000	1.90	860,437	0.14	0.07	167.23
2000	13.92	0.53	0.17	0.06	440 000	1.07	0.13	0.05	440 000	1.64	897,474	0.09	0.01	9.05
2010	14.26	0.54	0.16	0.07	727 000	1.30	0.06	0.01	727 000	1.22	649,841	0.06	0.01	8.61

[a] 5% critical value is 0.11.
[b] The 5% critical values for sample size of 142 is about 4.57 as given in Table 1 of Urzúa (2000).

Table 2
Distribution parameterization for India.

Year	Lognormal				Pareto				General Pareto					
	$\hat{\mu}$	$\hat{\sigma}^2$	KS[a]	MSE	$\hat{x}_{\min}$	$\hat{\alpha}^P$	KS[a]	MSE	$\hat{x}_{\min}$	$\hat{\phi}$	$\hat{\theta}$	KS[a]	MSE	LM[b]
1950	12.29	1.08	0.12	0.03	16,000	0.38	0.32	2.27	16,000	2.90	735,665	0.13	0.06	152.95
1960	12.68	0.83	0.13	0.05	36,000	0.45	0.32	1.42	36,000	2.36	735,811	0.14	0.10	88.20
1970	13.08	0.72	0.17	0.06	98,000	0.62	0.28	0.47	98,000	1.90	690,254	0.15	0.06	40.57
1980	13.48	0.66	0.18	0.08	209,000	0.80	0.22	0.16	209,000	1.54	650,964	0.14	0.04	18.07
1990	13.83	0.64	0.23	0.09	299,000	0.81	0.25	0.16	299,000	1.22	625,210	0.20	0.06	18.84
2000	14.12	0.64	0.20	0.10	523,000	1.03	0.14	0.03	523,000	1.06	551,980	0.14	0.02	3.29
2010	14.37	0.65	0.20	0.12	746,000	1.16	0.12	0.01	746,000	1.03	600,111	0.13	0.02	1.42

[a] 5% critical value is 0.18.
[b] The 5% critical values for sample size of 58 is about 4.49 Table 1 of Urzúa (2000).

This recording system registers detailed information of a person including name, date of birth, parents, and residential area (Pines et al., 1998). This record keeping was extensively utilized not only for identification, but also to control population mobility, particularly from rural to urban areas to ensure structural stability (Macleod, 2001). As a result, between the 1950s and 1970s, China was primarily a rural economy with about 83% of the population inhabiting in agrarian communities, and urban migration was stagnant with only 17% of the population residing in urban areas. Consequently, government regulation prevented cities from their natural growth, defying Gibrat's Law of proportionate growth of cities. These underpinnings are reflected by our findings that the Chinese city size in the earlier periods did not follow Zipf's law, rather adhered to lognormal.

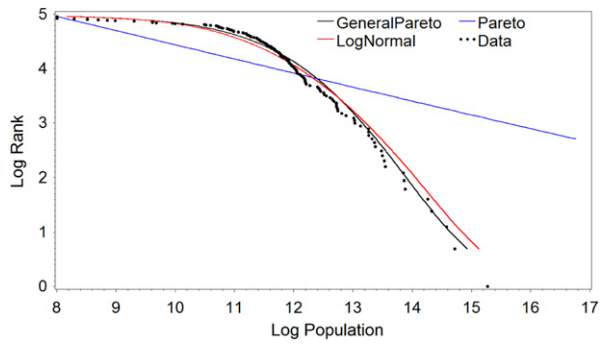
Performance of the Pareto distribution is a reversal of that of lognormal (refer to Table 1), in that it fits the Chinese city size data poorly from 1950 to 1990 and predicts better for the recent decades (2000 and 2010). The *KS* statistics are significantly above the critical values for the periods 1950–1990, but well below the critical value in 2010. These results are also supported by the *MSE*s and Zipf plots (Fig. 1(a)–(g)). For the decades 1950–1990, the estimate of the Pareto exponent is well below one, ranging from 0.26 to 0.50. But in the last two decades the Pareto exponent approaches one, but never becomes Zipf based on the *LM* statistics presented below. These results for the last two decades could be, as elaborated below, indicative of Chinese economic reforms which allowed for migration from rural to urban areas.

The general Pareto spans the lognormal in the early decades and Pareto distribution in more recent decades. The results show that the general Pareto fits the city size distribution in China more accurately for all seven decades, as reflected by the *KS* statistics which are below the 5% critical value (except for 1990), the small *MSE*s, and the Zipf plots (Fig. 1(a)–(g)). The results of the Lagrange multiplier test shows that Zipf's law is strongly rejected for every decade at the 5% significant level of 4.57, even though the values of the *LM* statistics tend to decrease steadily from 1950 to 2010. This result and the plots demonstrate that the city size distribution is approaching Zipf's law, but does not quite reach Zipf yet.

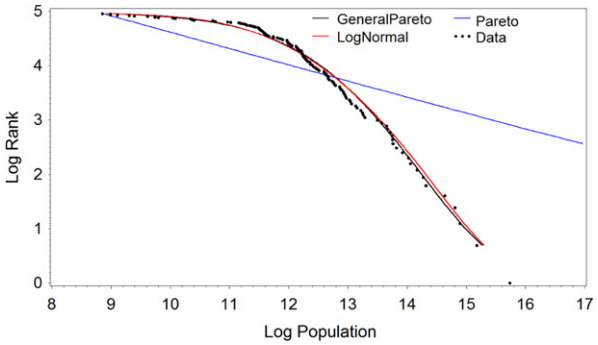
In the late 1970s the Chinese government implemented two major policies: economic reforms in 1978 and a population control policy of one child per family in 1979. The first policy was to augment the economic growth to alleviate poverty, and the second policy was to improve social, economic, and environmental problems. The economic reform spurred growth in industrial areas and increased demand for workers in urban cities. The government, realizing the Hukou registry is an impediment to economic development and importance of labor in manufacturing sectors, began to gradually, but not completely, relax the migration restriction from villages to cities (Wang, 2008). Thus, migration to urban areas took its roots originating from the economic reforms in the late 1970s. Since this policy was not fully liberalized, the city size did not follow Zipf (or even Pareto) in the early part of the reform in 1980s and 1990s.

The one-child policy was not followed uniformly and had many exemptions. One such exemption was to allow rural families to have a second child if the first child is a girl. However, this policy was effectively followed in urban cities with a high compliance rate. Li (1995) found that in urban cities 91% of the mothers had only one child, whereas in rural areas only 59% of mothers had one child because of greater resistance to this policy. Consequently, the fertility ratio was 2.4 for all of China but only 1.3 for urban areas (Snyder, 2000). Thus, policy could have prevented urban cities from following its natural growth process and becoming Pareto in the 1980s and 1990s.

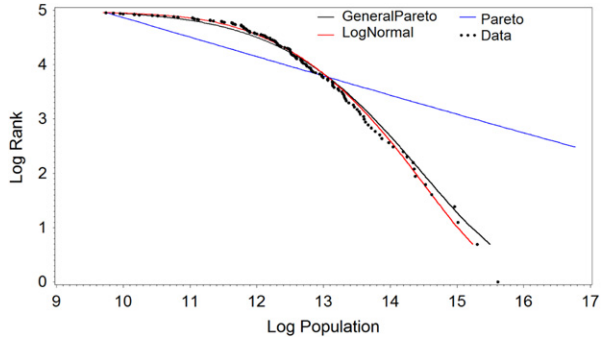
But as the economic reform and development accelerated, the government further relaxed the Hukou system in the mid-1990s and early 2000s and urban migration also gathered momentum. As a result, the growth process of cities tend to progress, albeit slowly, toward its natural process in recent years. Consequently, more than 50% of the population is living in urban areas since 2011 (The World Bank, 2014). Our empirical findings indeed underscore this change as evident from the city size distribution converging toward Pareto in the recent two decades (2000 and 2010). But, it has not become Zipf because the one-child policy likely slowed the natural growth process. During the third plenary session of the 18th Central Committee in 2013, Chinese Premier Li Keqiang put forth policy for a major overhaul of Hukou to further augment urban growth (Marshall, 2013). In addition, the one-child policy is also being relaxed. The revisions of these two policies will accelerate migration to urban areas which will cause the upper-tail city size distribution to continue to converge to Zipf.



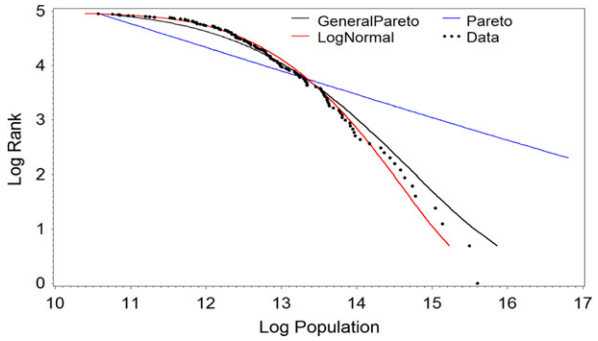
(a) 1950.



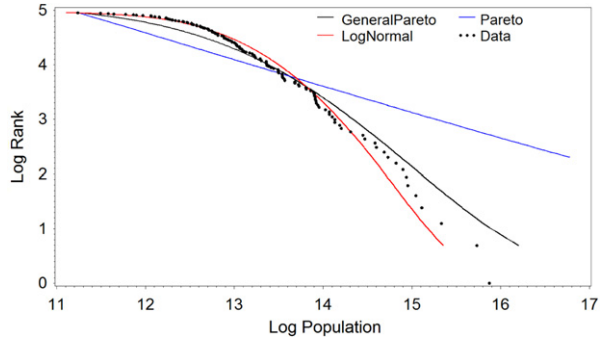
(b) 1960.



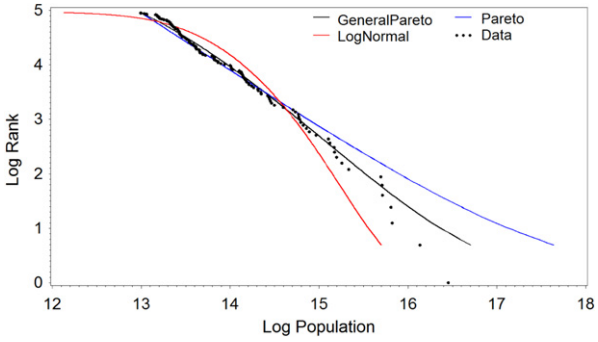
(c) 1970.



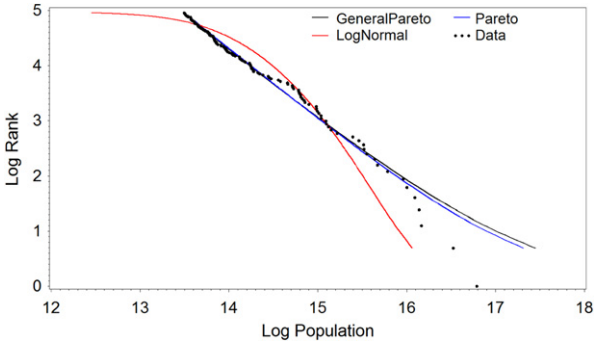
(d) 1980.



(e) 1990.



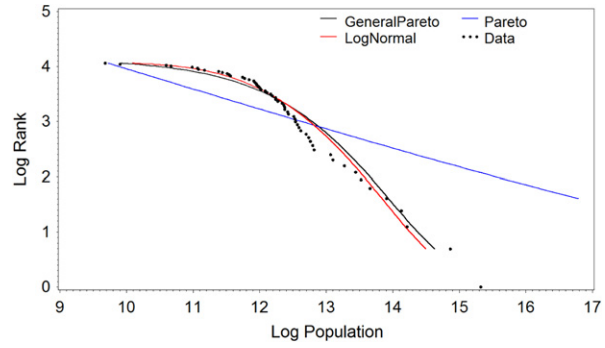
(f) 2000.



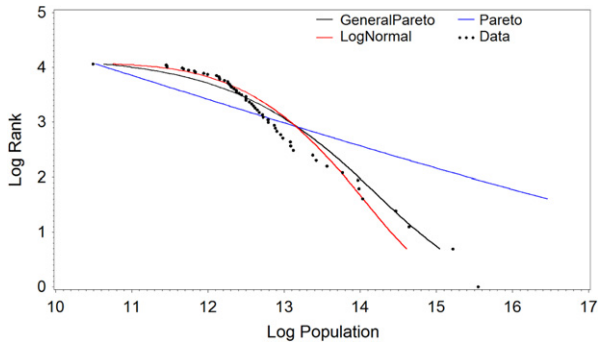
(g) 2010.

Fig. 1. Chinese population distributions.

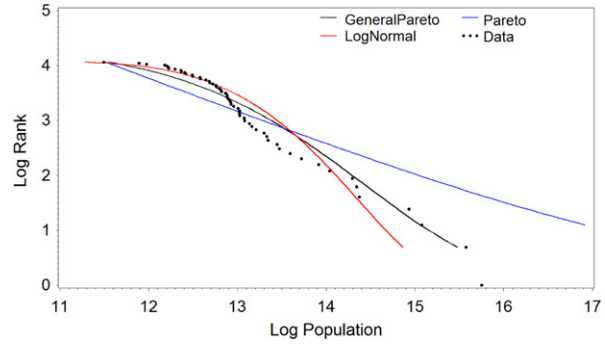
The estimated results for India reveal that the city size distribution is lognormal (refer to Table 2) from 1950 to 1980 as the *KS* statistics are at or below the 5% of 0.18, and as revealed by the low *MSE*s (0.03 and 0.08) and Zipf plots (Fig. 2(a)–(c)). During the first four decades of the sample period (1950s–1980s), India was largely an agrarian economy with more than 80% of the population living in the rural area (The World Bank, 2014). With dismal industrial development in these periods due to the license Raj economy, there was no economic incentive for the rural mass to migrate to urban areas because of the failure of the manufacturing sector to



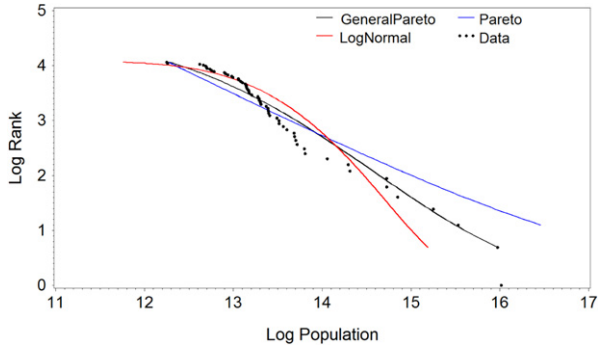
(a) 1950.



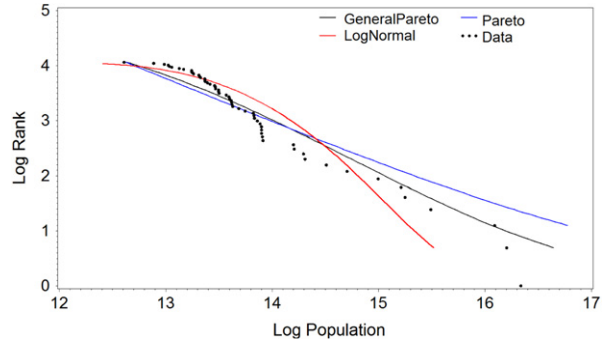
(b) 1960.



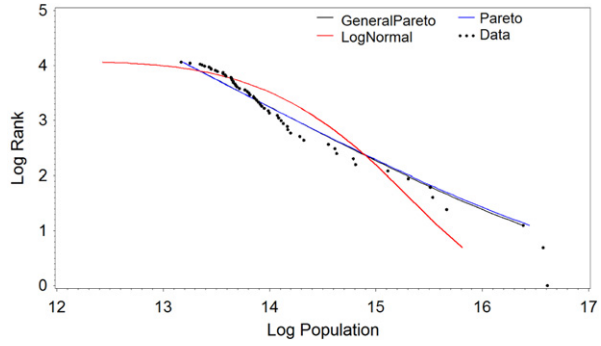
(c) 1970.



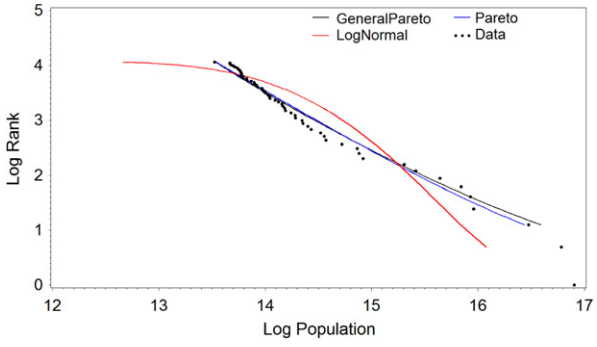
(d) 1980.



(e) 1990.



(f) 2000.



(g) 2010.

Fig. 2. Indian population distributions.

generate employment opportunities. Consequently, mobility from villages to cities was limited (Binswanger-Mkhize, 2012). However, unlike China, India is a democratic country and no legislative policy prevented migration to urban areas, as evident from percentage of urban dwellers showed a modest increase from 19% to 24% from 1960s to 1980s (The World Bank, 2014). As a result, city size distribution, stemming from natural migration, slowly and steadily approached Zipf from 1950s to 1990s. This convergence is borne out by the coefficient estimates of Pareto ($\hat{\alpha}^P$) and General Pareto ($\hat{\phi}$), both of which approach to one from the 1950s (refer

to Table 2). These results are corroborated by the calculated values of the *LM* statistics which steadily decrease from 152.95 in 1950 to 18.84 in 1990. The *KS* statistics gradually decrease for the Pareto distribution and are less than the critical value for general Pareto (except for 1990). This result is also supported by the *MSE*s which tend toward zero for these two distributions. The Zipf plots (Fig. 2(a) through (e)) for the period 1950–1990 also exhibit this trend.

The lognormal does not fit the city size distribution for 1990–2010. The *KS* statistics are greater than the 5% critical value and the *MSE*s are also larger (Table 2). The Pareto and general Pareto fit the data well, as determined by the *KS* statistics, which are below the 5% critical value for the last two decades, and also supported by the *MSE*s being closer to zero. The parameter estimates of Pareto and general Pareto distribution are closer to one. Also observe that general Pareto is flexible and mimics lognormal in the earlier periods and nests Pareto in the later period, and it consistently does better than lognormal or Pareto.

It is worth observing that the city sizes are Zipf for India in 2000 and 2010, as shown by the *LM* test, which fails to reject the null hypothesis at the 5% significant level of 4.49. Fig. 2(f) and (g) also illustrate this result. Gangopadhyay and Basu (2009) also find Zipf for Indian cities based on the *KS* test, but not the *LM* test which is more rigorous.

India began its economic reforms in the early 1990s which spurred economic growth, particularly in the industrial sector (Panagariya and Rajan, 2004). With this economic development, demand for workers in urban areas increased, which was accompanied by steady and slow migration from rural to urban areas. Consequently, city population experienced a more natural growth process, which resulted in the size distribution becoming Zipf.

4. Conclusion

This study shows that the largest cities in the two most populous countries in the world have similar trends: city size distribution is lognormal in the early periods and Pareto in 2010. However, as indicated by the Lagrange multiplier test, the city size distribution becomes the well-known Zipf for India for 2000 and 2010, but not for China. These results are consistent with the cross-country findings of Soo (2007), who reject Zipf for 30 of the 73 countries analyzed using the Hill (maximum likelihood) estimator.

Acknowledgment

The authors gratefully acknowledge an anonymous reviewer for providing valuable suggestions.

References

Anderson, G., Ge, Y., 2005. The size distribution of Chinese cities. Reg. Sci. Urban Econ. 35 (6), 756–776.
Binswanger-Mkhize, H.P., 2012. India 1960–2010: structural change, the rural non-farm sector, and the prospects for agriculture. Technical Report. Center on Food Security and the Environment, Stanford University.
Gabaix, X., 1999. Zipf's law for cities: an explanation. Quart. J. Econ. 114 (3), 739–767.
Gangopadhyay, K., Basu, B., 2009. City size distributions for India and China. Physica A 388 (13), 2682–2688.
Ioannides, Y.M., Overman, H.G., 2003. Zipf's law for cities: an empirical examination. Reg. Sci. Urban Econ. 33 (2), 127–137.
Krugman, P., 1996. The Self-Organizing Economy. Blackwell Publishers, Cambridge, Massachusetts.
Li, J.I., 1995. China's One-child policy: How and how well has it worked? A case study of Hebei Province, 1979–88. Popul. Dev. Rev. 21 (3), 563–585.
Macleod, C., 2001. China reviews 'apartheid' for 900m peasants. The Independent, June 10.
Marshall, J., 2013. China: urbanization and hukou reform. The Dipolmat, October 11.
Panagariya, A., Rajan, R., 2004. India in the 1980s and 1990s: a triumph of reforms. Wp/04/43, International Monetary Fund.
Pines, D., Sadka, E., Zilcha, I., 1998. Topics in Public Economics: Theoretical and Applied Analysis. Cambridge University Press.
Rosen, K.T., Resnick, M., 1980. The size distribution of cities: an examination of the pareto law and primacy. J. Urban Econ. 8 (2), 165–186.
Snyder, M., 2000. Govermental control and cultural adaptation: a comparison between rural and urban reactions to China's fertility control policies. Luce Foundation, Reed College. http://www.reed.edu/luce/documents/SnyderLuceReport.pdf.
Soo, K.T., 2007. Zipf's law and urban growth in Malaysia. Urban Stud. 44 (1), 1–14.
Stanley, M.H., Buldyrev, S.V., Havlin, S., Mantegna, R.N., Salinger, M.A., Eugene Stanley, H., 1995. Zipf plots and the size distribution of firms. Econom. Lett. 49 (4), 453–457.
The World Bank, 2014. World development indicators database. http://data.worldbank.org/data-catalog/world-development-indicators.
United Nations, 2011. World population prospects: the 2010 revision and world urbanization prospects: the 2011 revision. Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat, esa.un.org.
Urzúa, C.M., 2000. A simple and efficient test for Zipf's law. Econom. Lett. 66 (3), 257–260.
Wang, D., 2008. Rural–urban migration and policy responses in China: challenges and options. Working Paper No. 15, ILO Asian Regional Programme on Governance of Labour Migration.
Zipf, G.K., 1949. Human Behavior and the Principle of Least Effort. Addison-Wesley Press.